Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica e Tecnologie Informatiche


Relazione Finale

# Analisi Multivariata dei profili Facebook per l'uso problematico

Relatore Prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Correlatore Dott.ssa Claudia Marino
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Laureando Simone Righetto
Matricola N 1011441

Anno Accademico 2015/2016

# Indice

# Capitolo 1

# Introduzione

Questo elaborato è la realizzazione del lavoro svolto sotto la supervisione del prof.Finos, nell'ambito di un'indagine condotta dalla dott.ssa Marino del Dipartimento di Psicologia dello Sviluppo e della Socializzazione l'indagine riguarda l'uso dei social network degli studenti.Gli studenti che hanno deciso di partecipare all'indagine hanno infatti dovuto rispondere ad un questionario fatto dalla dott.ssa Claudia Marino e successivamente scaricare e consegnarLe i relativi dati Facebook, l'indagine infatti consisterà nell'estrarre i dati degli studenti e creare con essi un Dataset per poi procedere ad un'Analisi delle Correlazioni Canoniche (CCA) con il Dataset creato dalle risposte ricevute nei questionari e quindi vedere se esistono delle correlazioni tra i due Dataset.

## 1.1 Facebook

Facebook è un social network che vanta quasi 1,4 miliardi di utenti al mondo. Questo social network permette infatti di registrarsi gratuitamente e successivamente di interagire con tutti gli utenti presenti nella piattaforma.Ovviamente, nel corso degli anni Facebook ha avuto notevoli evoluzioni: si è passati dal semplice social dedicato a ritrovare persone che con il trascorrere del tempo si erano ormai perse o a stringere nuove amicizie con utenti già presenti nel social Network ad una nuova realtà virtuale che permette ora di creare eventi, pagine, gruppi,giochi, postare foto, scrivere commenti, condividere post e tanto altro ancora. Questo implica necessariamente la presenza,

di una notevole quantità di dati al suo interno e quindi di informazioni. Perciò obiettivo di questo lavoro è, oltre alla possibilità di unire i dati ricavati tramite questionario a quelli direttamente derivanti dal sito , ricavare delle descrizioni delle varie tipologie di utenti di facebook da poter confrontare con qualche profilo psicologico emerso. Inoltre, dati gli script che sono stati scritti per la raccolta e pulizia dei dati si è voluto creare una libreria per la raccolta dei dati personali e altresi fornire una descrizione analizzando il proprio profilo(myFBr).

## 1.2 Fasi preliminari

Nelle fasi iniziali del lavoro si è dedicato tempo ad analizzare il linguaggio con il quale risultano scritti i dati estratti da Facebook.Ciò che interessava capire era come fossero strutturati tali dati e come Facebook li rilasciasse una volta scaricati. Una volta capito che questi dati erano di tipo HTML(Hypertext Markup Language)[1]si è cercato di capire se esistesse una libreria R per poter gestire alberi e nodi, struttura del linguaggio HTML.Questa libreria esisteva già e si chiama XML(eXtensible Markup Language)[2] essa infatti permette di poter maneggiare come meglio si preferisce i nostri dati che sono appunto codificati con il linguaggio HTML. Utilizzando il linguaggio XPATH, che permette di individuare con precisione i nodi all interno di un documento XML, è stato possibile estrapolare diversi dati relativi all uso personale di facebook di ogni singolo profilo in maniera adeguata. In particolare, sono stati raccolti dati quantitativi relativi alle amicizie, agli eventi, alle foto, ai messaggi privati e alle attività espresse sulla bacheca. Il Dipartimento di Psicologia dello Sviluppo e della Socializzazione ha fornito i dati di 113 studenti dell'omonimo corso di studi.Per rendere il tutto più veloce ed automatico si è deciso quindi di creare funzioni in R che permettessero di entrare nei profili Facebook ed estrarre le informazioni da noi cercate.

[1]* Hypertext Markup Language, linguaggio di base per la creazione di pagine web che utilizza delimitatori, in gergo chiamati "tag", che "contengono" le informazioni.

[2]** eXtensible Markup Language, linguaggio che permette la strutturazione di informazioni sfruttando dei "tag".

## 1.3 Struttura dati del profilo

I dati dei profili Facebook a disposizione si sono rivelati una miniera ricca di informazioni, infatti avere accesso a tutte queste informazioni non una cosa facile in quanto si è dovuto chiedere agli studenti di poter entrare nel loro profilo per scaricare i loro dati( il percorso per scaricare i dati da facebook è log in con il proprio account Impostazioni->Generali->Scarica un copia dei propri dati Facebook). Una volta ottenuti tutti questi dati si è potuto notare quante informazioni ci fossero al loro interno: informazioni riguardati il numero di accessi, amicizie, messaggi, foto, post in bacheca ecc..(in Fig. 1.1 avete un'immagine di come sono strutturati i dati all'interno del file .zip). Non tutti i file sono risultati ugualmente facili da manipolare con R, in alcuni non vi erano alcune sezioni forse perché l'utente che li aveva forniti li aveva anche manomessi oppure per la presenza di errori nella fase di download o nella fase di copia e incolla da un dispositivo all'altro; ne consegue che non tutti e 113 profili sono stati usati per il nostro esperimento.Alla fine siamo riusciti ad arrivare ad un totale di 108 unità. C'è da notare che per estrarre e creare il dataset contenenti tutte le informazioni, le funzioni create non sono state facili e intuitive da creare come si potrebbe essere erroneamente indotti a pensare, ma questo lo vedremo nel capito successivo.

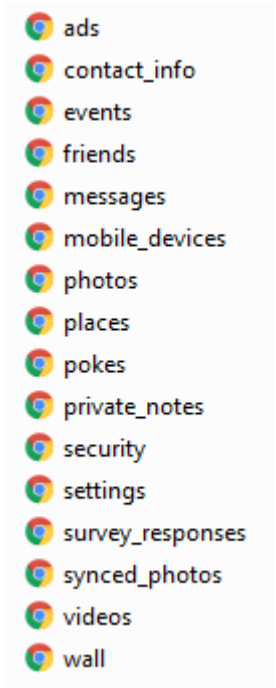


**Figura 1.1:** Dati all'intero del file .zip

# Capitolo 2

# Estrazione ed analisi dati Facebook

## 2.1 Estrazione dei dati

Durante la fase di estrazione dei dati, come detto in precedenza, è stata curata la tipologia dei dati in possesso. Una volta capito di quale tipologia essi fossero, si sono create delle funzioni in R in grado di entrare nell'albero dei nostri file HTML e in grado di percorrerlo fino ad arrivare al nodo di interesse(nella tabella 3.1 si ha un esempio delle funzioni presenti nel pacchetto myFBr). Vediamo ora nel dettaglio alcune di queste funzioni cosi da poter mostrare come i dati siano stati estratti. Come primo esempio si prenda in questione la funzione numberFB che restituisce in una riga le seguenti informazioni: sesso, n. di sessioni e di accessi, n. di amici di varia categoria, n. di post, n. di attività del wall delle varie categorie, n. di eventi con varia risposta, n. di foto, n. di posti e n. di messaggi privati.

```
numberFB <- function(percorso,dateS=TRUE,dateE=TRUE){
  sesso=getGender(percorso)
  if(sesso=="Uomo"){sesso=1}
  else{sesso=0}
  accessi=getAccessSessions(percorso,dateS,dateE)
  amici=getFriends(percorso)
  nPost=getPost(percorso,dateS,dateE)
  nEventi=getEvents(percorso,dateS,dateE)
```

```
  nFoto=getPhotos(percorso,dateS,dateE)
  mess=getMessages(percorso,dateS,dateE)
  wall=getWall(percorso,dateS,dateE)
  nPosti=getPlaces(percorso,dateS,dateE)
  dataReg=getRegDate(percorso)
  dati<-cbind(sesso,accessi,amici,nPost,wall,nEventi,nFoto,nPosti,mess,dataReg)
  return(dati)
}
```

Questa funzione richiama a se tante altre funzioni e per capire meglio come queste operino entriamo nel dettaglio della funzione getFriends. Come si può vedere, è una funzione che richiede una variabile chiamata «percorso»impostata precedenza con all'interno il percorso del nostro file(Es C: \Utenti \...).La funzione quindi incollerà il nostro percorso ad \html \friend \... e lo inserirà in perA. Successivamente, essa andrà a leggere l'intero file e grazie al pacchetto XML utilizzerà un xPath per entrare nello specifico nodo ( ad esempio per accet è ul[1]), al cui interno ci sarà l'informazione desiderata. Una volta ottenute tutte le informazioni, si costruirà il dataset amici che ritornerà una matrice con una riga e quattro colonne che avranno i rispettivi nomi (accet,richi,ricev,rimos) .

```
getFriends <- function(percorso){
  perA=paste(percorso,"/html/friends.htm", sep="")
  pg=htmlParse(perA)
    accet=length(getNodeSet(pg,"//div[@class='contents']/div/ul[1]/li/text()"))
  richi=length(getNodeSet(pg,"//div[@class='contents']/div/ul[2]/li/text()"))
  ricev=length(getNodeSet(pg,"//div[@class='contents']/div/ul[3]/li/text()"))
  rimos=length(getNodeSet(pg,"//div[@class='contents']/div/ul[4]/li/text()"))
    "amici" <- structure(.Data = list(accet,richi,ricev,rimos),
                         names = c("accettati", "richEff", "richRic","rimossi"),
                         row.names = c(1:1),
                         class = "data.frame")
  return(amici)
}
```

Nella funzione getPost, invece si può notare un elemento nuovo: questa funzione va a leggere un'informazione vincolando una data di partenza, allo scopo di sapere quanti post il profilo ha pubblicato in bacheca, e di dare dei limiti temporali specificando una data di inizio dateS e/o una di fine dataE. Così facendo, è possibile limitare la raccolta dei dati ad un periodo a piacere. Questa funzione agisce in maniera analoga all' esempio appena esposto (senza date di riferimento), ma in aggiunta effettua un controllo sulle date,e quindi fornisce il numero di post pubblicati successivamente alla data di inizio e/o precedentemente alla data di fine indicate. Se non vengono indicati i parametri delle date, la funzione procede ugualmente calcolando tutte le informazioni che riesce a reperire nel file fornito.

```
getPost <- function(percorso,dateS=TRUE,dateE=TRUE){
  perW=paste(percorso,"/html/wall.htm", sep="")
  pg=htmlParse(perW)
  atti=getNodeSet(pg,"//div[@class='meta']/text()")
  n=length(atti)
  post=0

  if(dateS==TRUE && dateE==TRUE){
    post=length(getNodeSet(pg,"//div[@class='meta']"));
  }else if(dateS==TRUE){
    if(n>0){
      for(i in 1:n){
        temp=inDateIT(as.character(.estraielemento(atti[[i]])))
        if(temp<=dateE){
          post=post+1
        }
      }
    }
  }else if(dateE==TRUE){
    if(n>0){
      for(i in 1:n){
        temp=inDateIT(as.character(.estraielemento(atti[[i]])))
        if(temp>=dateS){
```

```
          post=post+1
        }
      }
    }
  }else{
    if(n>0){
      for(i in 1:n){
        temp=inDateIT(as.character(.estraielemento(atti[[i]])))
        if(temp>=dateS && temp<=dateE){
          post=post+1
        }
      }
    }
  }

return(post)
}
```

| Funzione | Descrizione |
|---|---|
| getFriends | ritorna i dati relativi al numero di amici |
| getPhotos | ritorna i dati relativi al numero di foto |
| getMessage | ritorna i dati relativi al numero di messaggi |
| getPost | ritorna i dati relativi al numero di post totali |
| getPlaces | ritorna i dati relativi al numero di posti condivisi |
| getGender | ritorna un dato relativo al proprio sesso |

**Tabella 2.1:** Alcune funzioni all'interno del pacchetto myFbr

## 2.2 Analisi dati

Una volta creato il nostro Dataset che sarà composto dalle seguenti variabili:

- "sesso" = genere del profilo
- "nSessioni" = numero di sessioni aperte, cioè numero di dispositivi differenti
- "nAccessi" = numero di accessi effettuati
- "amiciAcce" = numero di richieste di amicizia accettate
- "RichEff" = numero di richieste di amicizia effettuate (senza risposta)
- "RichRic" = numero di richieste di amicizia ricevute (senza risposta)
- "AmiciRim" = numero di richieste di amici rimossi
- "WnPost" = numero di post (presenti in bacheca)
- "WAmici" = numero di amicizie strette (presenti in bacheca)
- "WNStato" = numero di stati (presenti in bacheca)
- "WNPiace" = numero di piace (presenti in bacheca)
- "WNLike" = numero di like (presenti in bacheca)
- "WNCondivisi" = numero di condivisioni (presenti in bacheca)

- "WNLink" = numero di link (presenti in bacheca)
- "WNGiocato" = numero di volte che si è giocato (presenti in bacheca)
- "eventiConf" = numero eventi confermati
- "eventiForse" = numero eventi forse
- "eventiRiff"= numero eventi rifiutati
- "eventiNoRisp" = numero eventi senza risposta
- "NFoto" = numero di foto caricate
- "NPosti" = numero di posti deve ci si è segnalati
- "messaggi" = messaggi privati mandati e ricevuti
- "dateReg" = date di registrazione del profilo

Si vada ora a vedere alcune analisi di tipo descrittivo fatte sui dati Facebook.

```
   accettati         richEff          richRic          rimossi
 Min.   :  21.0   Min.   :  1.00   Min.   :  0.00   Min.   :   0.00
 1st Qu.: 421.0   1st Qu.:  3.00   1st Qu.:  4.00   1st Qu.:   0.00
 Median : 621.0   Median :  7.00   Median : 16.00   Median :   9.00
 Mean   : 729.7   Mean   : 12.48   Mean   : 44.42   Mean   :  65.88
 3rd Qu.: 912.0   3rd Qu.: 15.00   3rd Qu.: 42.00   3rd Qu.:  59.00
 Max.   :4752.0   Max.   :242.00   Max.   :784.00   Max.   :1085.00
    amicizia          stato            piace          condiviso
 Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
 1st Qu.:  48.0   1st Qu.:  4.00   1st Qu.: 22.00   1st Qu.:  0.0
 Median :  85.0   Median : 18.00   Median : 50.00   Median :  1.0
 Mean   : 109.8   Mean   : 38.37   Mean   : 82.87   Mean   : 11.7
 3rd Qu.: 113.0   3rd Qu.: 38.37   3rd Qu.: 84.00   3rd Qu.:  9.0
 Max.   :2043.0   Max.   :680.00   Max.   :931.00   Max.   :459.0
  linkAltri         linkTuo         partecipato     postInBacheca
 Min.   : 0.000   Min.   :  0.000   Min.   :  0.00   Min.   :  0.00
 1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:  3.00   1st Qu.: 38.00
 Median : 1.000   Median :  3.000   Median : 10.00   Median : 68.06
 Mean   : 1.873   Mean   :  9.611   Mean   : 20.72   Mean   : 68.06
 3rd Qu.: 2.000   3rd Qu.:  9.610   3rd Qu.: 20.72   3rd Qu.: 90.00
 Max.   :37.000   Max.   :378.000   Max.   :254.00   Max.   :233.00
 aggiornImgProf     nuovaFoto         postTotali       eventiConf
 Min.   : 0.000   Min.   :   0.00   Min.   :   0.0   Min.   :  0.00
 1st Qu.: 0.000   1st Qu.:   2.00   1st Qu.: 222.0   1st Qu.:  3.00
 Median : 2.000   Median :  15.00   Median : 389.0   Median :  8.00
 Mean   : 3.146   Mean   :  59.48   Mean   : 577.4   Mean   : 18.11
 3rd Qu.: 4.000   3rd Qu.:  59.48   3rd Qu.: 600.0   3rd Qu.: 19.00
 Max.   :29.000   Max.   :1137.00   Max.   :5118.0   Max.   :217.00
```

```
  eventiRif       eventiNoRisp        nPhoto             mess
 Min.   :  0.00   Min.   :   0.0   Min.   :   0.00   Min.   :     0
 1st Qu.:  0.00   1st Qu.:  48.0   1st Qu.:   0.00   1st Qu.:   875
 Median :  1.00   Median : 102.0   Median :   0.00   Median :  2094
 Mean   : 11.19   Mean   : 144.4   Mean   :  46.93   Mean   :  5790
 3rd Qu.:  6.00   3rd Qu.: 183.0   3rd Qu.:  21.00   3rd Qu.:  4492
 Max.   :468.00   Max.   :1784.0   Max.   :1116.00   Max.   :205546
```

In Fig.2.1 si ha un grafico che mette in evidenza le correlazioni delle nostre variabiabili, si può notare da questo grafico come le variabili eventiConf e partecipato siano molto correlate tra loro ed in maniera positiva difatti l'elisse è molto sottile e rivolto verso destra ,nella Fig 2.2 si può notare che la correlazione tra le due variabili è pari a 0.92. Lo stesso ragionamento lo si può fare anche per le due variabili richEff e richRic, solo che in questo caso si ha una correlazione quasi nulla, infatti si ha un cerchio bianco, ed in Fig.2.2 si può notare che la correlazione che c'è tra le due variabili è pari a 0.01.

Dopo le analisi preliminari di tipo descrittivo, nel prossimo capitolo si andrà ad approfondire l'argomento delle CCA allo scopo di vedere se esiste qualche correlazione tra i questionari e i dati Facebook rilasciati dagli studenti del Dipartimento di Psicologia.

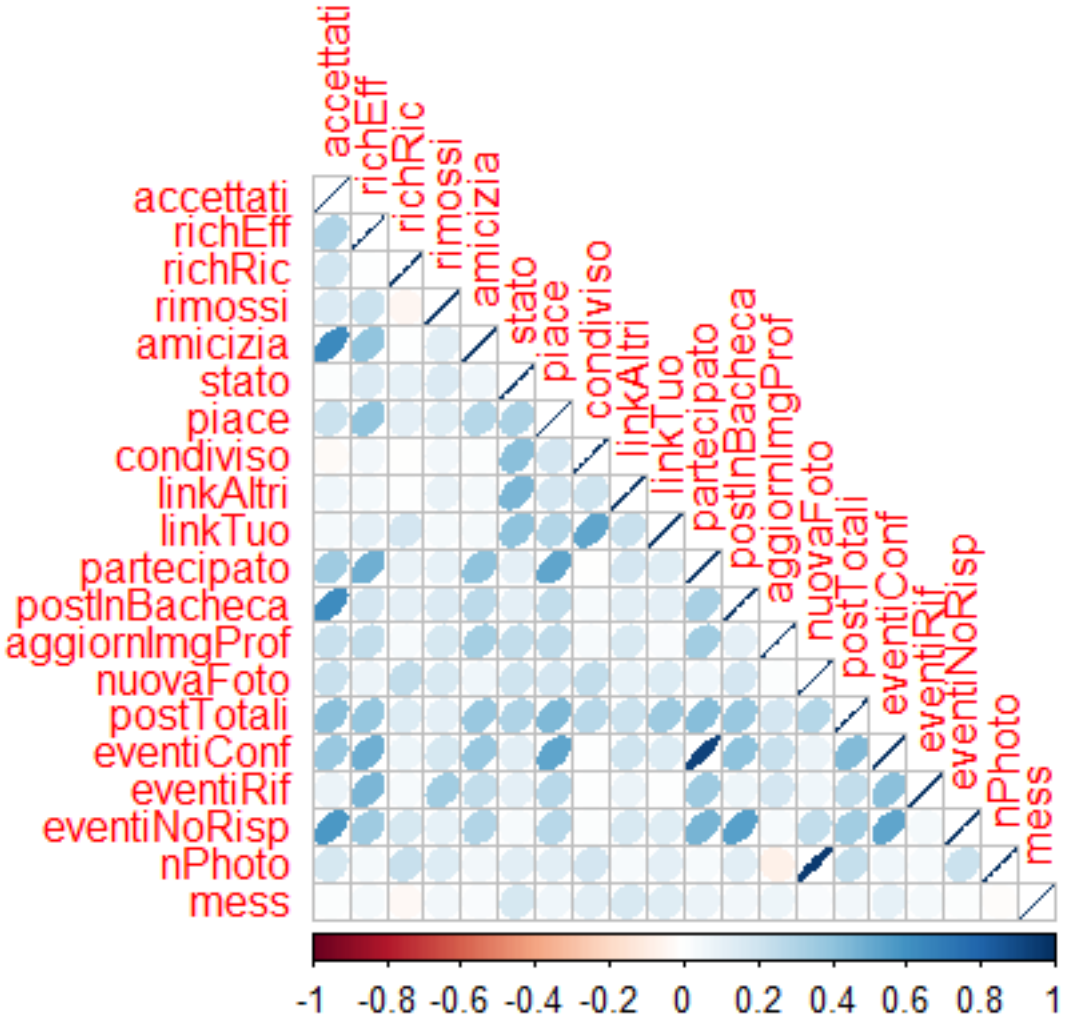


**Figura 2.1:** Grafico delle correlazioni delle variabili dei dati Facebook

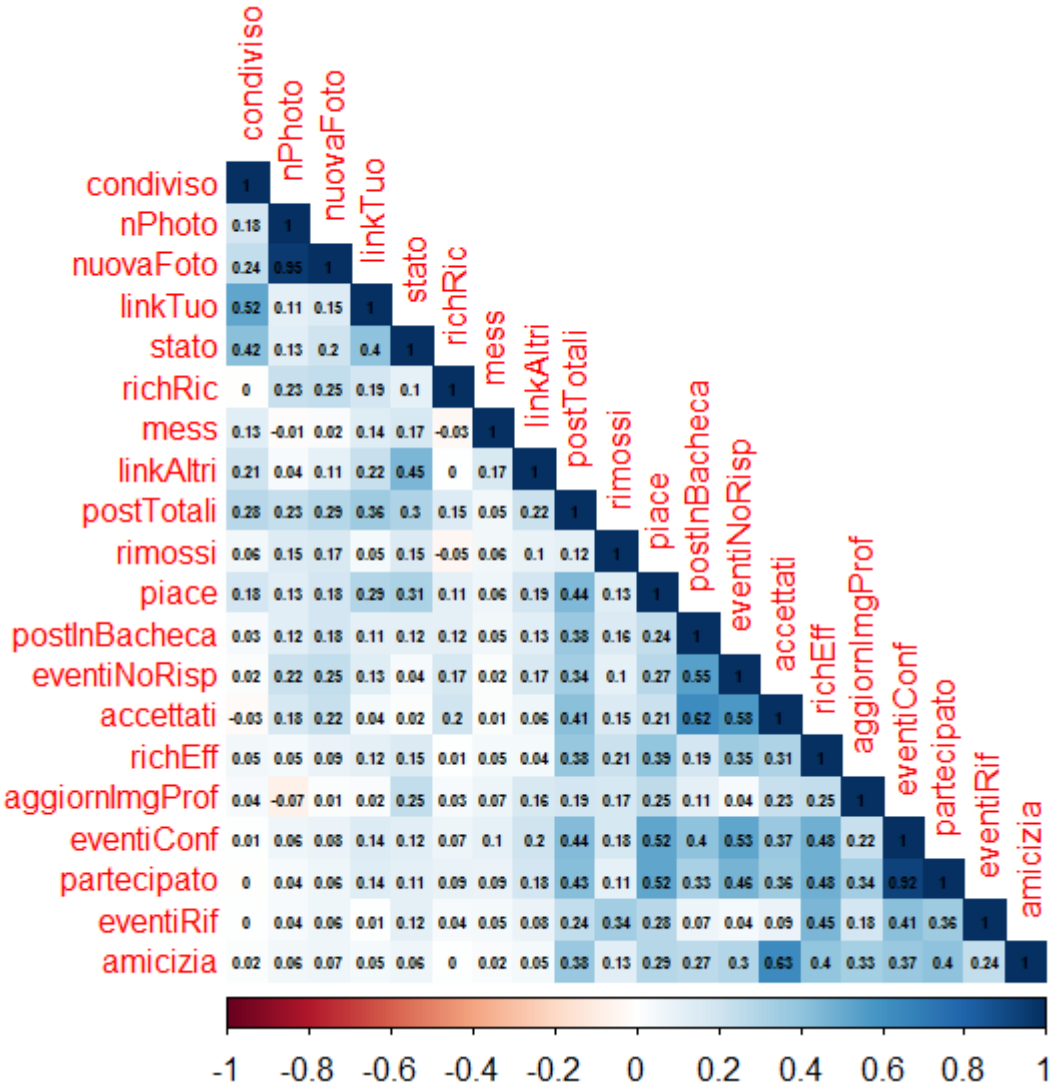


**Figura 2.2:** Grafico delle correlazioni delle variabili dei dati Facebook

# Capitolo 3

# I test GPIUS2 e IMQ-A

## 3.1 Introduzione ai Test

Una volta estratti tutti i dati e dopo aver creato il dataset con tutti i dati Facebook degli studenti, con la dott.ssa Claudia Marino si è lavorato per creare il Dataset dei test Psicologici, inizialmente si sono realizzati diversi incontri per discutere e stabilire quelli che erano i punti chiave di questi test.Ovviamente da un lato si è aiutato Claudia a cercare e spiegare come creare un dataset in R, nell'altro la dott.ssa invece ha spiegato quali fossero i punti principali dei test e che cosa questi volessero significare, cosi da non fare errori gravi (come l'omissione di qualche item importante), non dare la giusta importanza ai test e quindi incorrere in errori di mal interpretazione.

## 3.2 Test Generalized Problematic Internet Use Scale 2 (GPIUS2)

Il test GPIUS2 è un riadattamento fatto dalla dott.ssa Marino per l'utilizzo di Facebook della Caplan (2010). Theory and measurement of generalized problematic Internet use: A two-step Approach. Computers in Human Behavior(GPIUS2)[2]. Il riadattamento di Facebook è costituito da domande che hanno come obbiettivo quello di capire quale sia l'utilizzo che l'utente fa di Facebook,vale a dire, mirato ad osservare se l'utente ha una dipendenza da

Facebook. Le domande infatti sono mirate a capire se, per esempio, l'utente preferisce Facebook per interagire con le persone, se invece abbia delle dipendenze dal social network(in questo caso, nei momenti in cui egli è offline sente il desiderio di collegarsi per vedere se è successo qualcosa di interessante) o se ancora ne fa un uso esagerato(situazione in cui lo stesso utente passa tanto tempo collegato) .

Questo test dunque ha l'obbiettivo di capire se l'utente possa essere classificato come utente a rischio dalla dipendenza da Facebook.

```
Formula test GPIUS2
#dataset$F : scala posi
lista.items=paste("F", c(4, 10, 15), sep="")
dataset$posi=crea.scala(dataset[,lista.items])
dataset$posi

#dataset$F : scala mood_reg
lista.items=paste("F", c(1, 6, 12), sep="")
dataset$mood=crea.scala(dataset[,lista.items])
dataset$mood

#dataset$F : scala cognitive preoccupation
lista.items=paste("F", c(2, 12, 14), sep="")
dataset$cognitive=crea.scala(dataset[,lista.items])
dataset$cognitive

#dataset$F : scala compulsive use
lista.items=paste("F", c(3,9,7), sep="")
dataset$compulsive=crea.scala(dataset[,lista.items])
dataset$compulsive

#dataset$F : scala negative outcomes
lista.items=paste("F", c(5, 13, 8), sep="")
dataset$negative=crea.scala(dataset[,lista.items])
dataset$negative
```

Si vada a vedere ora una breve descrizione di tutte le scale che compongono il nostro test Test GPIUS2:

- "Fposi"= la tendenza di un utente a preferire il social per l'interazione sociale con altre persone
- "Fmood"= indica che l'utente tende ad usare il social quando è demoralizzato o si sente solo
- "Fcognitive"= indica che l'utente non si sente a suo agio, ansia, dovuta al fatto che è da troppo tempo disconnesso dal social network
- "Fcompulsive"=indica che l'utente fa un uso ossessivo di Facebook
- "Fnegative"= indica se Facebook ha influenzato negativamente la vita dell'utente , poichè preferisce il social all'uscire di casa

Ecco una tabella che mostra i primi dieci valori del nostro test:

| Fposi | Fmood | Fcognitive | Fcompulsive | Fnegative | Ftotal |
|---|---|---|---|---|---|
| 3.000000 | 3.000000 | 1.666667 | 1.333333 | 1.000000 | 2.000000 |
| 1.000000 | 1.333333 | 1.333333 | 1.333333 | 1.000000 | 1.200000 |
| 1.000000 | 3.000000 | 3.666667 | 2.333333 | 1.000000 | 2.000000 |
| 1.333333 | 4.000000 | 2.333333 | 4.333333 | 1.333333 | 2.466667 |
| 1.333333 | 3.000000 | 2.000000 | 2.333333 | 1.000000 | 2.000000 |
| 1.000000 | 2.333333 | 3.333333 | 2.000000 | 1.000000 | 1.800000 |
| 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| 1.000000 | 1.666667 | 1.333333 | 1.000000 | 1.000000 | 1.200000 |
| 1.333333 | 2.333333 | 2.000000 | 4.000000 | 1.333333 | 2.266667 |
| 1.000000 | 1.333333 | 1.000000 | 1.000000 | 1.000000 | 1.066667 |

**Figura 3.1:** Tabella dei primi dieci valori del test GPIUS2

### 3.2.1 Statistiche descrittive

Alcune statistiche e grafici del test:

```
 Fposi           Fmood           Fcognitive      Fcompulsive
Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:1.00   1st Qu.:1.330   1st Qu.:1.000   1st Qu.:1.330
Median :1.00   Median :2.000   Median :1.670   Median :2.000
Mean   :1.57   Mean   :2.235   Mean   :1.873   Mean   :2.308
3rd Qu.:1.67   3rd Qu.:2.835   3rd Qu.:2.330   3rd Qu.:3.000
Max.   :8.00   Max.   :6.330   Max.   :7.000   Max.   :7.330
 Fnegative
Min.   :1.000
1st Qu.:1.000
Median :1.000
Mean   :1.323
3rd Qu.:1.330
Max.   :5.670
```

In Fig.3.2 si ha una tabella che illustra la correlazione tra i nostri items. Nella parte alta della tabella si può notare il valore della correlazione tra i vari items, ad esempio tra Fmood e Fnegative è pari a 0.55, mentre nella parte inferiore della tabella si ha un indice come siano correlati i vari items, ad esempio tra Fmood e Fcognitive si ha una correlazione positiva pari a 0.6.


**Figura 3.2:** Grafico delle correlazioni degli items del test GPIUS2

## 3.3 Test Internet Motive Questionnaire for Adolescents (IMQ-A)

Il test IMQ-A invece è anch'esso un riadattamento realizzato dalla dott.ssa Marino per l'utilizzo di Facebook della Bischof-Kastner et al. (2014) Identifying Problematic Internet Users: Development and Validation of the Internet Motive Questionnaire for Adolescents(IMQ-A)[1].Questo test ha lo scopo di verificare quali siano gli obbiettivi dell'uso di Facebook: le domande sono mirate a capire quali siano i motivi che spingono l'utente ad usare tale social,ad esempio se l'utilizzo è determinato da un abbattimento morale,da una volontà di non essere escluso, o di dimenticare i tuoi problemi,o ancora, da un sentimento di felicità legato allo stare connesso o di appartenere ad un determinato gruppo di amici. Questo test ha l'obiettivo di determinare il motivo per cui tale utente decide di usare Facebook.

```
Furmula test IMQ-A
#coping
lista.items=paste("E", c(1, 3, 5, 13), sep="")
dataset$coping=crea.scala(dataset[,lista.items])
dataset$coping

#conformity
lista.items=paste("E", c(2, 10, 15, 16), sep="")
dataset$conformity=crea.scala(dataset[,lista.items])
dataset$conformity

#enhancement
lista.items=paste("E", c(6, 7, 8, 14), sep="")
dataset$enhancement=crea.scala(dataset[,lista.items])
dataset$enhancement

#social
```

```
lista.items=paste("E", c(4, 9, 11, 12), sep="")
dataset$social=crea.scala(dataset[,lista.items])
dataset$social
```

Si vada a vedere ora una breve descrizione di tutte le scale che compongono il nostro test Test IMQ-A:

- "Ecoping"=indica che l'utente utilizza Facebook quando è demoralizzato, o per dimenticare le proprie preoccupazioni

- "Eenhancement"=l'utente utilizza il social network per divertirsi,o provare emozioni positive

- "Econformity"=l'utente fa uso di Facebook perché è una moda o per cercare di integrarsi ad un nuovo gruppo di amici

- "Esocial"=l'utilizzo che l'utente fa del social è per contattare, stringere nuove amicizie,o chattare con altri utenti

Ecco una tabella che mostra i primi dieci valori del nostro test:

| Ecoping | Eenhancement | Econformity | Esocial |
|---|---|---|---|
| 1.00 | 1.25 | 1.75 | 2.25 |
| 1.00 | 1.75 | 1.25 | 2.75 |
| 1.50 | 1.50 | 1.25 | 2.25 |
| 1.25 | 2.25 | 1.75 | 4.25 |
| 1.75 | 2.75 | 1.75 | 4.25 |
| 1.00 | 1.00 | 1.50 | 2.75 |
| 1.00 | 1.00 | 1.00 | 1.25 |
| 1.25 | 1.50 | 1.00 | 2.00 |
| 1.50 | 2.50 | 1.50 | 3.75 |
| 1.00 | 1.25 | 1.00 | 1.75 |

**Figura 3.3:** Tabella primi dieci valori del test IMQ-A

### 3.3.1 Statistiche descrittive

Alcune statistiche descrittive del test:

```
 Ecoping       Eenhancement   Econformity       Esocial
Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
1st Qu.:1.000   1st Qu.:1.25   1st Qu.:1.000   1st Qu.:2.000
Median :1.250   Median :1.75   Median :1.250   Median :2.250
Mean   :1.579   Mean   :1.73   Mean   :1.466   Mean   :2.487
3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:1.750   3rd Qu.:3.000
Max.   :4.250   Max.   :3.75   Max.   :3.750   Max.   :4.500
```

Nella Fig.3.4 si ha una tabella che mostra le correlazioni tra gli items del test IMQ-A. Anche in questa tabella come in quella del test GPIUS2, possiamo vedere come sono correlati i vari items tra loro e il valore delle correlazioni.

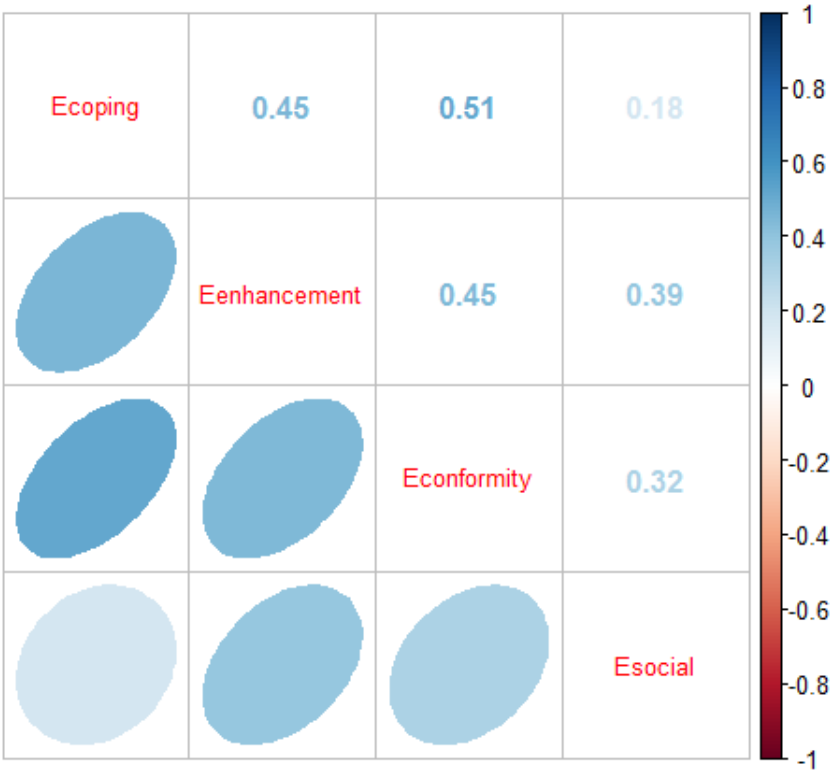


**Figura 3.4:** Grafico delle correlazione degli items del test IMQ-A

## 3.4 Gestione dei dati mancanti

In merito agli ostacoli incontrati nel corso della creazione dei suddetti test si vuole dedicare un'attenzione particolare alla gestione dei valori mancati(NA).Nell'ambito della statistica, questo argomento risulta spesso delicato, la soluzione più immediata che intuitivamente verrebbe in mente alla maggior parte delle persone è di andare ad eliminare i dati mancanti; questa però è una soluzione che risulta inappropriata: un dato mancante non ha valore zero e non è semplicemente inesistente,esso risulta invece mancante, ovvero da qualche parte esiste solo che non lo si possiede e questo implica, necessariamente, una diversa soluzione. Il metodo che si è voluto utilizzare per la soluzione degli NA è quello di rimpiazzarli invece con la media dei dati da noi conosciuti(non si parla di tutti i dati in possesso ma solo di quelli attribuiti

alla nostra variabile di interesse ad esempio se risulta esserci un NA nella variabile altezza si prenderanno in considerazione solo i dati riguardanti l'altezza e non anche quelli del peso). Per far ciò si è dovuto creare una funzione in R.Questo metodo è generalmente ritenuto un buon metodo, ma ha anche i suoi limiti: lo si consiglia solo se si hanno a disposizione almeno 80% dei dati,poichè l'utilizzo di una minore percentuale di dati rischierebbe di compromettere l'esperimento in quanto essi non basterebbero a spiegare la vera informazione, ma ne produrrebbero solo una forma distorta o incompleta di essa.

# Capitolo 4

# Analisi delle Correlazioni Canoniche CCA

## 4.1 Una breve introduzione alle Componenti Principali PCA

Il contesto in cui ci si muove è quello dell'analisi statistica di dati multivariati, cioè il caso in cui si ha un campione di n individui per ciascuno dei quali si sono osservate $p > 2$ caratteristiche (le variabili). Le componenti principali servono a riassumere un campione relativo a p variabili attraverso h variabili che vengono, appunto, chiamate componenti principali e sono combinazioni lineari delle variabili di partenza non correlate tra di loro e con $h < p$. In questo modo si tenta di risolvere a un tempo e il problema della dimensionalità del campione e quello della multicollinearità (Everitt and Dunn [2001][3]). Le componenti principali possono essere il risultato finale dell'analisi o possono essere un risultato intermedio: ad esempio nella regressione multipla può essere opportuno, anziché usare le variabili esplicative a disposizione direttamente, ottenere le componenti principali e usare queste ultime come esplicative questo a un tempo riduce la dimensionalità del problema e fornisce esplicative non correlate. Le c.p., d'altra parte, possono anche essere viste come risultato finale di un'analisi in quanto riassumono le caratteristiche delle unità: ad esempio, si supponga di disporre, per n soggetti, dei

risultati in p diverse prove e di voler stilare una graduatoria dei soggetti, la prima delle c.p. è la sintesi dei risultati che garantisce la maggiore distinzione tra gli n individui. La derivazione delle componenti principali avviene sequenzialmente: si cerca una combinazione lineare della variabili di partenza massimizzandone la varianza, poi si cerca una seconda combinazione lineare che massimizzi la varianza e che sia incorrelata con la precedente e cosi via. Le variabili cosi costruite hanno importanza decrescente, la prima essendo quella che discrimina di più le osservazioni. L'importanza può essere anche misurata dalla percentuale di varianza che ciascuna di esse spiega. Quindi la differenza sostanziale che hanno le PCA dalle CCA è che quest'ultime studiano le interazioni tra due diversi insiemi di variabili anziché entro uno stesso insieme.Dove però una sarà la matrice dei predittori mentre l'altra sarà la matrice predetta( un esempio nella Fig. 4.1 dove si vedono le due matrici di dati X).

$$Z_{n\times(p+q)} = \begin{array}{c} \begin{matrix} \text{PREDETTE} & & & \text{PREDITTRICI} & & \end{matrix} \\ \begin{pmatrix} y_{11} & \cdots & y_{1p} & x_{11} & \cdots & x_{1q} \\ y_{21} & \cdots & y_{2p} & x_{21} & \cdots & x_{2q} \\ \vdots & & & \vdots & & \\ y_{n1} & \cdots & y_{np} & x_{n1} & \cdots & x_{nq} \end{pmatrix} \end{array} \quad \text{MATRICE CAMPIONARIA DEI DATI}$$

**Figura 4.1:** La matrice Z formata da X e Y

## 4.2 Analisi delle Correlazione Canoniche

Prima di iniziare con le CCA si voleva precisare che il numero di unità per l'esperimento sono aumentate passando da 108 a 341.L'esperimento condotto sembrerebbe infatti aver avuto un grande successo e grazie ad un passaparola generale si sono ottenute ulteriori adesioni al progetto qui illustrato.Si veda dunque un esempio semplice ma efficace per capire meglio come leggere i risultati delle correlazioni canoniche. Si prendano come esempio i dati riguardanti le quantità annuali di pesce venduto al mercato Ittico di Chioggia nel periodo 1945-2008.

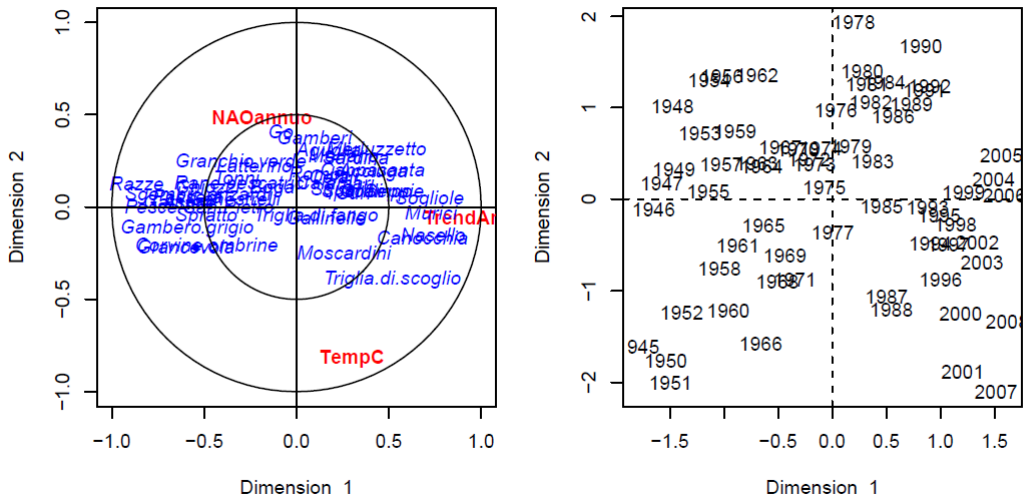


**Figura 4.2:** Grafico delle CCA del mercato Ittico di Chioggia

Si analizzi l'immagine che, come detto è il risultato finale delle CCA della vendita di pesce del mercato Ittico di Chioggia (1945-2008). Come prima cosa si notano le due componenti principali:la prima componente nell'asse X la seconda componente nell'asse Y ai lati del grafico; al centro invece si vedono i dati di due colori diversi: blu le tipologie di pesce e rosso le variabili indice NAO (North Atlantic Oscillation) annuo, il trend annuale e la temperatura.Le ultime tre variabili costituiranno i predittori dell'esperimento.Osservando innanzitutto le componenti principali si può notare dal grafico il Trend Annuo domina la prima componente mentre la TempC domina la seconda.Si può inoltre notare anche che il Go e i Gamberi sono molto vicini all' indice NAO e questo indica che le due variabili sono correlate tra loro e perciò all'aumentare dell' indice NAO aumenteranno anche i Gamberi e i Go, invece la Triglia di Scoglio è collocata dalla parte opposta all indice NAO, quindi all'aumentare dell'indice NAO diminuirà la Triglia di Scoglio. Questo ragionamento si può fare per tutte le variabili predittrici. Dopo questa breve introduzione su come interpretare questo modello di grafico si torni all'argomento d'interesse principale.Innanzitutto, si deve scaricare ed installare all'interno di R la libreria CCA.Fatto ciò si vada ad usare la funzione"cc(datiFB,GPIUS2)", per eseguire le CCA tra i dati Facebook e la GPIUS2, idem per la IMQ-A.Si

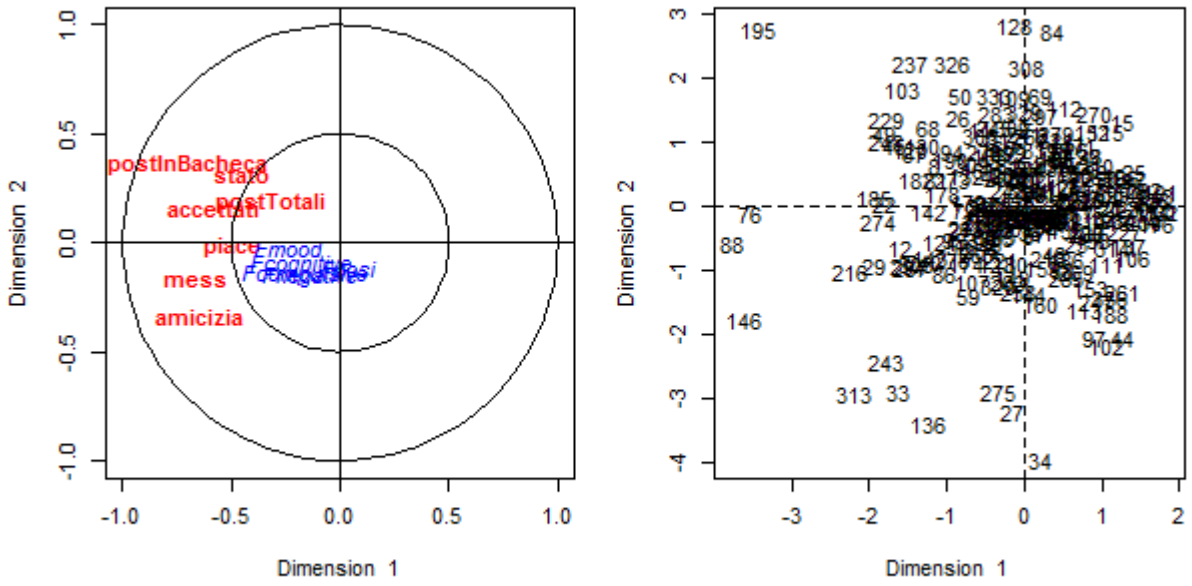


**Figura 4.3:** Grafico delle CCA tra gli items del test GPIUS2 e le variabili di Facebook

osservino ora i due grafici prodotte dalle CCA, Fig.5.3 e 5.4.

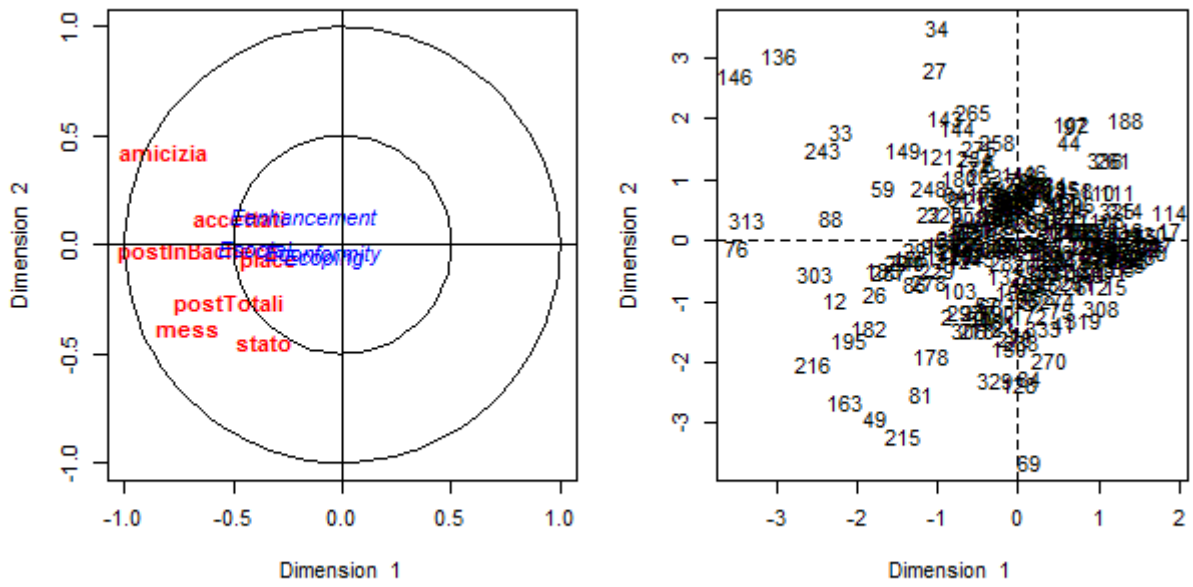


**Figura 4.4:** Grafico delle CCA tra gli items del test IMQ-A e le variabili di Facebook

## 4.3 Conclusioni

Come si può notare dai grafici sovra illustrati, i due test di colore blu sono tra loro molto vicine e questo indica che i loro indici sono molto correlati tra loro. La stessa cosa non si può invece dire per tutte le variabili dei dati Facebook: esse in effetti non risultano essere interamente correlate tra loro infatti, il primo grafico, che prende in considerazione il test GPIUS2, mostra come le variabili: mess, piace e accettati siano dei buoni predittori per tale scala.Mentre, per quanto riguarda il test IMQ-A le variabili: accettati, piace e PostinBacheca sembrano essere dei buoni predittori per la seconda scala.Dal grafico del test IMQ-A possiamo notare come le scale accettati, Eenhencement ed Esocial siano particolarmente vicine,non a caso, Esocial è la scala legata all'uso dei social per stringere nuove amicizie, o per entrare in un gruppo di amici, a sua volta la variabile Eenhencement è invece la scala che indica se l'utente fa uso di Facebook per divertirsi, o per provare nuove emozioni.Tra le tre variabili risulta dunque una forte correlazione, e la cosa tutto sommato non ci sorprende cosi tanto, è difatti del tutto lecito che al crescere di amicizie accettate il valore della scala Esocial si alta e che quindi un utente che utilizza Facebook per stringere nuove amicizie abbia un valore elevato di amicizie accettate.

# Bibliografia


[1] Christina Bischof-Kastner, Emmanuel Kuntsche e Jorg Wolstein. "Identifying Problematic Internet Users: Development and Validation of the Internet Motive Questionnaire for Adolescents (IMQ-A)". In: *J Med Internet Res* 16.10 (ott. 2014), e230. DOI: 10.2196/jmir.3398. URL: http://www.ncbi.nlm.nih.gov/pubmed/25299174.

[2] Scott E. Caplan. "Theory and Measurement of Generalized Problematic Internet Use: A Two-step Approach". In: *Comput. Hum. Behav.* 26.5 (set. 2010), pp. 1089–1097. ISSN: 0747-5632. DOI: 10.1016/j.chb.2010.03.012. URL: http://dx.doi.org/10.1016/j.chb.2010.03.012.

[3] Brian S Dunn Everitt, Graham Brian S Everitt e Graham Dunn. *Applied multivariate data analysis*. Rapp. tecn. 1991.